Australian Government
Department of Defence
Defence Science and
Technology Organisation

A New Interpretation of the Shannon Entropy Measure

Lewis Warren

National Security and ISR Division
Defence Science and Technology Organisation

DSTO-TN-1395

ABSTRACT

Although more than sixty years have elapsed since Shannon's seminal information entropy paper the literature reveals that there are divergent opinions of what it actually measures. From its similarity to Boltzmann entropy in statistical mechanics, the most common view is that it measures information disorder and uncertainty. Based on an inductive derivation of the expression we propose a new interpretation relating to the use of symbols to uniquely identify probabilistic messages. Applying this new meaning to Shannon's English language experiment we extract a new interpretation of those results. Moreover, the new understanding of Shannon entropy also has the potential to improve the effectiveness of intelligence analysis applications.

RELEASE LIMITATION

Approved for public release

Published by

National Security and ISR Division
DSTO Defence Science and Technology Organisation
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: 1300 333 363
Fax: (08) 7389 6567

AR-016-211
January 2015

APPROVED FOR PUBLIC RELEASE

A New Interpretation of the Shannon Information Entropy Measure

Executive Summary

Shannon stated in his seminal paper that any monotonic function of the number of messages generated by a source can be regarded as a measure of the information being generated. One consequence of using any function of the number of messages and their probabilities is that the meaning of the measure becomes obscure, and consequently is open to many different interpretations as evidenced by the literature on the topic. This report aims at clarifying what Shannon entropy actually measures. From first principles we show that Shannon entropy simply represents the average number of quantised symbols required to encode or identify an incoming probabilistic message. This follows directly from the definition of the logarithmic function. We also show that the logarithmic base represents the degree of symbol quantisation. Finally we apply the new interpretation to Shannon's English language experiment which yields a new meaning to those results. The primary motivation behind this re-examination of Shannon entropy is to gain a more precise interpretation so that the measure can be more effectively applied when analysing Defence information, such as for word sense disambiguation in automated text analysis.

Reference:

Shannon, C. (1948) A mathematical theory of communication, The Bell System Technical Journal, 27(3), pp. 379-423.

Contents

1. Introduction

The Shannon entropy measure [1,2] has been widely studied in the information theory literature. In the popular literature [3] it is frequently considered to be a measure of order and uncertainty. Many authors have proven that it is the only measure that satisfies a number of fundamental desiderata or axioms for an uncertainty measure. One view is that Shannon entropy represents the maximum amount by which the state space or message length can be compressed. Another view is that it represents the information gain by an event from a probabilistic space, sometimes also called the "expected surprise". One example of this view [4] states:

> *…the entropy of a system with prescribed mean energy can be interpreted as the maximum amount of missing information.*

One important justification of the Shannon entropy expression, which is often quoted in the literature, is that it is the only measure to satisfy the axioms that Shannon defined. But as we will see some authors also question the validity and completeness of those axioms. Perhaps some of the confusion can be attributed to the original circuitous and difficult to follow derivation by Shannon. So although the fundamental expression has been extended many ways [5][6] into so-called generalised entropy measures by applying conventional probability laws, as well as it being used in practical compression algorithms, we believe there still remains considerable confusion about its true meaning. Furthermore, several authors [4][7][8][9][10] have identified some implicit limitations of Shannon entropy which should be acknowledged. It is also noteworthy that there is still some confusion about the meaning of the older Boltzmann statistical mechanics type of entropy of physical systems, as indicated by the following comments [10], although "disorder" is the most common interpretation for that type of entropy:

> *The qualitative character of entropy has been discussed widely… The metaphoric images invoked for entropy include "disorder", "randomness", "smoothness", "dispersion", and "homogeneity". In a posthumous fragment, Gibbs mentioned "entropy as mixed-up-ness." Images such as these can be useful and important, but if taken too literally they can confuse as well as enlighten, and when misused they can result in simple error.*

To commence this analysis, the conclusions of a recent author are presented to set the context for this Shannon entropy analysis. Next, as a first step towards clarification, Hartley's original derivation of his information measure is presented since it is a special case of the Shannon expression. The meaning of Shannon entropy is then exposed by applying inductive logic to derive the expression. Following that, some computational examples are also presented to indicate the key behavioural aspects of Shannon entropy, some of which dispel some common beliefs. Finally, some implications of the suggested meaning will be presented.

2. Some Recent Comments on Shannon Entropy

Recently, in "The Mathematical Theory of Information", Kahre [11] proposes a new approach to the measurement of information and makes certain assertions about Shannon entropy (H):

The belief in entropy as the only adequate measure of information is deeply rooted. This can be seen in the information theory literature, where the uniqueness of H is proved in different ways using the same basic method. A set of axioms is proposed and declared to be necessary for an information measure, and then the "uniqueness theorem" is proved, i.e. that H is the only function satisfying the proposed axioms…These properties are rather strongly tailored to the Shannon entropy (Aczel and Daroczy 1975). Hence the proofs have a flavour of circularity. The goal seems to be to find the weakest possible set of axioms (Gottinger, 1975 p.7-8) sufficient to prove the uniqueness of H. [11, p.107]

For a function H to be a measure of the information content of a message set A, Shannon suggested [1] it to be "reasonable" for it to have the following properties:

1. It should be continuous in the $P(a_i)$, where P is a probability.
   i.e. small variations in P cause only small variations in the measure.

2. It increases monotonically with N if $P(a_i) = 1/N$
   i.e. with more events there is more information when an event occurs.

3. It satisfies the decomposition rule for joint entropy of two discrete random variables:`
   i.e. $H(A,B) = H(A) + H(B|A)$
   $= H(B) + H(A|B)$
   $= H(A) + H(B)$ if A and B are independent random variables.
   i.e. for a sequence of independent events the total information should be the sum of the partial entropy measures.

Some other comments by Kahre follow.

Even if we accepted the axioms, the uniqueness of H(A×B) as an information measure does not follow. The axioms only imply that H(A) is a unique measure of information content, but H(A) as information content is shared by other measures of information (A×B). e.g by the Bernoulii information measure.

And,

Shannon H is the largest amount of bits transmitted with 100% reliability. Hence the Shannon information measure is the number of sure bits. [11, p. 90]

And,

The axioms 1-3 cannot however, be accepted as fundamental properties of an information measure, because many important information measures are in conflict with the axioms. For instance, members of the utility family such as Gambler's gain, or reliability, violate the decomposition rule. [11, p.108]

And,

Thus Shannon H is not a measure of information, but rather an upper limit HB (Boltzmann entropy) of the true information HG (Gibbs Total Entropy) [11, p.220]

The previous comments by Kahre indicate that after 60 years there are still some members of the information theory community not altogether satisfied with the prevailing understandings of Shannon entropy. Needless to say, this does not necessarily diminish the importance of Shannon entropy within information theory.

3. Measures of Information or Uncertainty

Some distinctions between measures of uncertainty, measures of information, and functions of uncertainty will now be described.

3.1 Measures of uncertainty

A measure, by definition, quantifies a property which by its own definition is capable of being measured. A measure of uncertainty then requires a measurable definition of uncertainty. However, uncertainty itself is an ambiguous concept because there are many types of uncertainty, as evinced by several proposed taxonomies of uncertainty [12][13]. Consequently, the type of uncertainty must be clearly identified if a measure of uncertainty is to be developed. For example, an estimate of the degree of approximation, ambiguity, or variance may be used to define a measure interval for a variable such as: Length ±5%, Velocity ± 10%, or Mean value ± 2 standard deviations for a statistical sample.

Until the first half of the twentieth century event likelihood as measured by probability was the primary type of uncertainty being quantified This chance type of probability, also called aleatory probability, was usually estimated from evidential data. Another probability variant termed "subjective probability" is often used to estimate a single event's likelihood in the absence of data, e.g. an estimate of the probability that you will have a car accident tomorrow. This type is called "subjective" because the basis for the estimate is some sort of knowledge from experience, or a feeling residing inside an individual, and when such knowledge is minimal the estimate tends to a guess. In recent decades, there has also been an increasing focus on non-probabilistic uncertainty and several different measures have been proposed [5][14][15][16,17][18]. Lotfi Zadeh, the founder of fuzzy set theory, has also proposed [18] a Generalized Theory of Uncertainty stating:

…Uncertainty is an attribute of information. A fundamental premise of the Generalized Theory of Uncertainty is that information whatever its form, may be represented as what is called a generalized constraint. In the Generalized Theory of Uncertainty a probabilistic constraint is viewed as a special, albeit important, instance of a generalized constraint.

Nevertheless, there are still some gaps and shortfalls in non-probabilistic uncertainty modelling, especially with measures of total uncertainty in bodies of information where hybrid forms of uncertainty exist. These shortfalls concern how to identify different types of uncertainty, how to represent them for quantification, and finally how to derive a composite uncertainty estimate. In general, measures of uncertainty can be considered as macroscopic indicators of the degree of clarity and conciseness of data, i.e. as levels of vagueness and ambiguity.

3.2 Measures of information

Information is also a nebulous concept relating to both the meaning and content of a message, as well as to the number or amount of messages. And since interpreting the content of a message invokes semantic complications, the normal focus of measures of information is on the amount of information, which is often thought of as the complement of the degree of uncertainty. But while this general relationship may exist, it is not an exact complement for quantitative measures. The reason for that is that they both measure different concepts as will be subsequently explained. And similar to Zadeh's Generalized Theory of Uncertainty, Klir has also proposed a Generalized Information Theory [19] by which many different kinds of information can be represented using formalisms that address the different kinds of information. In general, measures of information can be considered as indicators of the degree of definition or organisation of event states and the complexity of data set elements. In the communications theory context addressed by Shannon, this then relates to the degree of difficulty in uniquely identifying an incoming message from the range of possible messages and superimposed noise.

3.3 Functions of uncertainty or information

In contrast to measures of uncertainty or information, which are directly related to the amount and types of input information, functions calculate values from expressions in which the uncertainty information about a variable is the input. Thus, a measure calculated by a function is an indirectly derived value based on some uncertainty representation, e.g. probability. Then the meaning of the dependent measure so computed can only be interpreted from the reason for the application of the particular function. So if sin(x) is used as an uncertainty measure, some relationship between the information characteristic x and the sin function is implied. Similarly, using $H = \log_b (x)$ necessarily implies $b^H = x$. Shannon stated [1] on the first page of his seminal 1948 paper that:
any monotonic function of the number of messages can be "regarded" as a measure of information, in lieu of the actual number of messages.

However, the difficulty with using any function is that the meaning of the measure becomes unclear, since it diverges from the number of messages which is the actual amount of information. Thus, a function generates a covariate to the quantity of uncertainty or information.

4. A Re-examination of Shannon Entropy

The Shannon entropy expression relates to the process of mapping messages from a source into coded symbol combinations to enable the identification of a message by a receiver. An important question is how to minimise the loss (as inability to identify a message) when using symbol combinations where only the probability of a range of possible messages is known. Using traditional parametric statistics this could be addressed by setting a confidence level and deriving limits based on the distribution mean and a number of standard deviations. However, the Shannon entropy expression reflects a different approach. Figure 1 illustrates the process of encoding messages from a probabilistic source using a fixed number N of quantised symbols S to communicate M messages to a receiver. Information loss can occur when there are an insufficient number of quantised symbol combinations, or codes, to identify an incoming message from the large number M emanating from the source. Since the degree of symbol quantisation determines the number N of symbols in a string of fixed length to encode the possible number of messages M, the aim is to select a number N such that the number of possible messages that cannot be identified is a minimum. For zero information loss, the number of quantised symbol combinations must be at least equal to the number of probable messages.

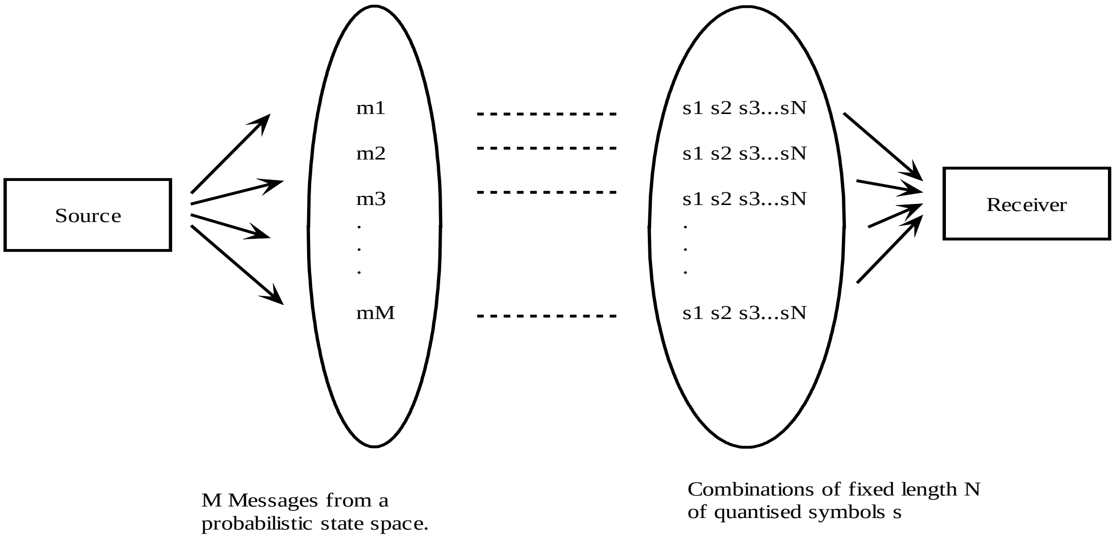
Source
m1
m2
m3
.
.
.
mM
s1 s2 s3...sN
s1 s2 s3...sN
s1 s2 s3...sN
.
.
.
s1 s2 s3...sN
Receiver
M Messages from a probabilistic state space.
Combinations of fixed length N of quantised symbols s

***Figure 1**: Communication of messages using symbols*

4.1 Hartley Information Derivation

We will now examine the derivation of a quantitative measure of "information" by Hartley published in his seminal paper [20] in 1928. This early measure is revisited because it is a special case of Shannon entropy, when all messages are completely random with equal probabilities of occurrence.

Hartley's Terminology:

S = Number of code symbols available as degree of quantisation (bits = 2, digits = 10)
n = Number of symbols in sequence
H = amount of information associated with n selections ("H" presumably for Hartley)

Itemised Steps in Hartley's Derivation

1) Fundamentally, an increase in n increases information content (H) so:

$$H \propto n \quad \text{or} \quad H = K\,n \qquad \text{(i)}$$

where K is a constant *which depends on S symbols available at each selection.*

2) Take two systems with different quantised symbols S1 and S2, and constants K1 and K2

3) Choose n1 and n2 for the two systems to yield an equal number of possible sequences (i.e. messages), then :

$$S1^{n1} = S2^{n2} \qquad \text{(ii)}$$

$$\text{and} \quad H1 = H2 \quad \text{(same number of sequences)}$$

$$\text{or} \quad K1\,n1 = K2\;n2 \qquad \text{(iii)}$$

4) Take Logs of (ii):

$$\log S1^{n1} = \log S2^{n2}$$

$$n1 \log S1 = n2 \log S2 \qquad \text{(iv)}$$

5) Substitute (iv) into (iii) for n1:

$$K1\left(\frac{n2 \log S2}{\log S1}\right) = K2\;n2$$

$$\frac{K1}{\log S1} = \frac{K2}{\log S2} \qquad \text{(v)}$$

6) (v) only holds if $Ki = K0 \log Si$, (vi)

where K0 is the same for all systems

7) Substitute (vi) into (i),

$$H = (K0 \log Si)\,n = K0\,n \log Si = K0 \log Si^{n} \qquad \text{(vii)}$$

8) *Since K0 is arbitrary, we may omit it if we make the logarithmic base arbitrary. The particular base selected fixes the size of the unit of information.*

9) Then, $H = \log Si^{n} = \log N$, (viii)

where $N = Si^{n}$ = Number of symbol sequences

The amount of information encoded in n selections of S quantised symbols is equal to S^{n}. So the Hartley measure of information is actually a log *function* of the amount of information, even though Hartley called it the "amount of information". The question then is what does this log function H really represent? This will be explored in the next section.

4.2 An inductive derivation of Shannon entropy

Shannon's entropy expression will now be derived, starting from a basic algebraic equation that relates the number of bits {0,1} required to identify a range of possible messages from a source which are completely random. Consider Figure 2 which illustrates the combinations of three bits available to identify or encode eight equally likely messages, events, states, or values (m1-m8). Figure 2 only depicts the combinations of symbols possible and should not be interpreted as a probability tree.

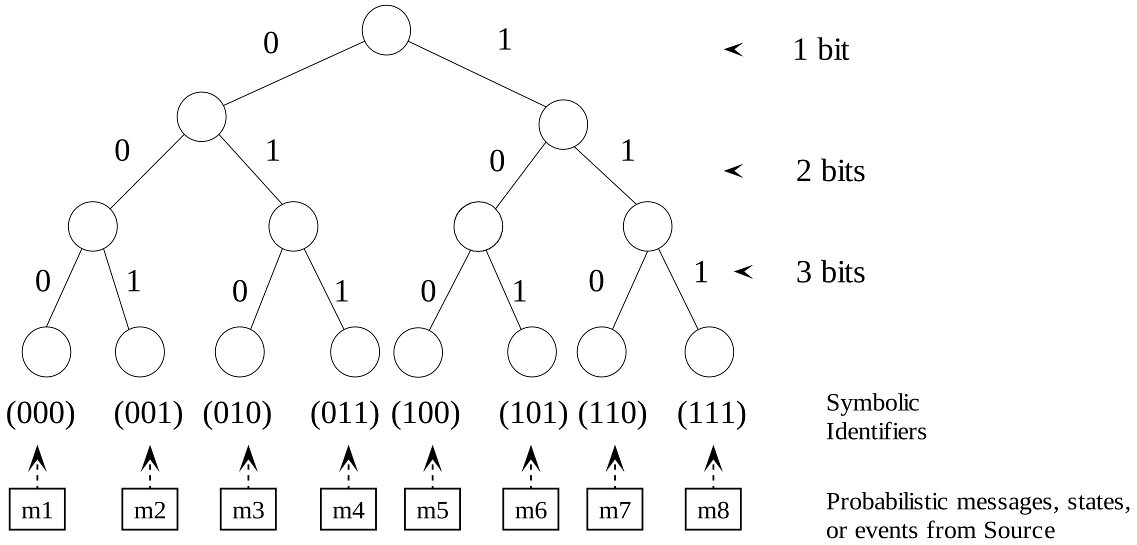

Figure 2: *Combinations of three binary symbols*

From Figure 2 we can define the equation:

$$\left(\frac{1}{2}\right)^3 = \frac{1}{8} \tag{1}$$

The fraction on the left side of equation (1) represents one of two states selected at each binary symbol, and the right side represents the identification of one message arriving from the possible range of messages.

Then with binary symbols {0,1} the following general expression defines the number of bits N required to identify a particular message, event, or state, from the number of messages A:

$$\left(\frac{1}{2}\right)^N = \frac{1}{A} \tag{2}$$

Then to represent eight messages using N decimal digits {0-9}:

$$\begin{aligned}\left(\frac{1}{10}\right)^N &= \frac{1}{8}\\ 10^N &= 8\\ \textit{and}\ \ N &= \log_{10} 8\\ &= 0.9\ \ \textit{decimal digits}\end{aligned} \tag{3}$$

Note that N need not be an integer and one decimal digit can identify 10 messages.

Generalising (3): $10^N = A$ and $N = \log_{10} A$ (4)

Expression (4) is the seminal Hartley Information Measure with log base 10.
Also, for the binary symbol space in Figure 2 and completely random messages:

$$\left(\frac{1}{2}\right)^3 = \frac{1}{8} = P_u\ ,\ \textit{where}\ P_u\ \textit{is the uniform probability for each message or state}$$

Then $2^3 = \frac{1}{P_u}$, $\quad \therefore\ 3 \ = -\log_2 P_u \ = \mathrm{N}$ (5)

Now (5), the number of bits N to identify or represent a message with a uniform probability of occurrence, can be generalised for messages with non-uniform probabilities because the expression need not be constrained to uniform probabilities. The reason for that is that N is a non-probabilistic variable itself, and only has a representational or encoding relationship to an event's probability of occurrence, as well as to whatever determines that probability.

Then, the number of bits $\mathrm{N_i}$ to represent a message i with probability $\mathrm{P_i}$ is,

$$N_i = -\log_2 P_i \tag{6}$$

which has been termed the Wiener entropy.

Then, by applying the standard statistical expectation expression the *expected* number of bits to represent a message arriving from the range of probable messages is:

$$\begin{aligned}&= \sum_{i=1}^{A} (\textit{Prob of message } i)\big(\textit{Number of Bits required for message } i\big)\\ &= \sum_{1}^{A} P_i\left(-\log_2 P_i\right)\\ &= -\sum_{1}^{A} P_i \log_2 P_i\ ,\ \ \textit{where A is the total number of probable messages.}\end{aligned} \tag{7}$$

The above is an inductive derivation of the discrete Shannon entropy expression, generalising from a simple example with uniform probabilities to a non-uniform probability event space with an expectation estimate for the number of bits to represent a message arriving from the source of probable messages.

Thus we can conclude:

1) Shannon entropy (H) is only relevant to probabilistic types of uncertainty as present in a probabilistic event space.
2) H represents the expected number of bits to identify a message (state, event, or value) from a probabilistic space of messages.
3) Because it is an average, some messages will require more bits than the entropy estimate.
4) H is *not* the minimum number of bits required to reliably transmit a probabilistic message, i.e. not the number of "sure" bits of Kahre.
5) H is *not* a quantification of uncertainty but a function of uncertainty.
6) H is an uncertainty covariate as a function dependent on the probability distribution of multiple possible messages (or events, states, or values).

4.3 Some illustrative examples

The following examples demonstrate how H increases with an increase in the number of messages, and/or some message probabilities are around the maximum uncertainty of p = 0.5. Figure 3 shows H as a function of *single* probability values. It should be noted that this distribution is skewed with the maximum not at p = 0.5, as it is only for *two* mutually exclusive binary messages with probabilities p and q, where q =1-p (as in Shannon's original paper).

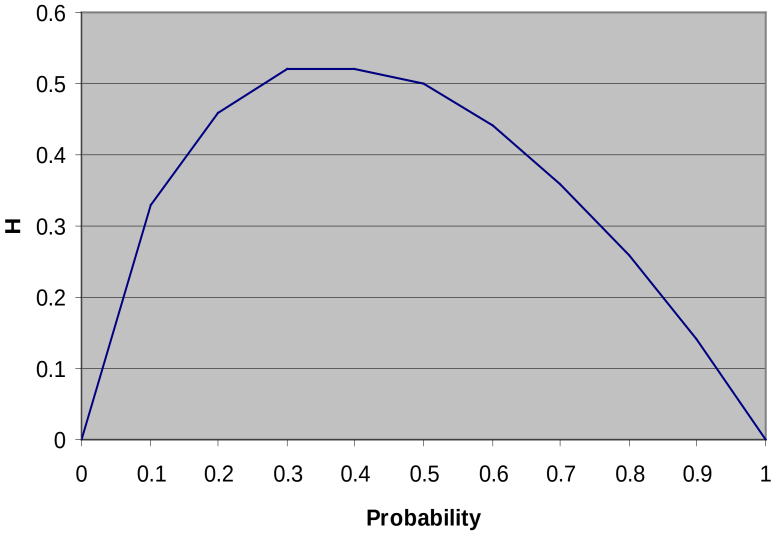

***Figure 3**: Shannon Entropy as a Function of a Single Probability*

5 Messages

1) With probabilities {4(0.05), 1(0.8)}

$$H = -[\,4(0.05)\log_2(0.05) + 0.8\log_2(0.8)\,]$$
$$= -[0.2\log_2(0.05) + 0.8\log_2(0.8)\,]$$
$$= 1.121 \text{ bits}$$

2) With probabilities {2(0.05), 0.1, 0.3, 0.5 }

$$H = -[2(0.05)\log_2(0.05) + 0.1\log_2(0.1) + 0.3\log_2(0.3) + 0.5\log_2(0.5)]$$
$$= 1.784 \text{ bits}$$

3) With uniform probabilities { 5(1/5) }

$$H = -5(1/5)\log_2(1/5) = \log_2(0.2) = 1.623 \text{ bits}$$

8 Messages

4) With uniform probabilities { 8(1/8) }

$$H = -8(1/8)\log_2(1/8) = 3.000 \text{ bits}$$

52 Messages

5) With uniform probabilities { 52(1/52) }

$$H` = -52(1/52)\log_2(1/52) = 5.700 \text{ bits}$$

6) With probabilities { 30(0.01), 20(0.02), 2(0.15) }

$$H = -[30(0.01)\log_2(0.01) + 20(0.02)\log_2(0.02) + 2(0.15)\log_2(0.15)] = 5.073 \text{ bits}$$

7) With probabilities { 50(0.01), 1 (0.4), 1 (0.1) }

$$H = -[50(0.01)\log_2(0.01) + (0.4)\log_2(0.4) + (0.1)\log_2(0.1)] = 8.943 \text{ bits}$$

From these illustrative calculations we can conclude:

- With a uniform distribution H increases from 5 to 8 to 52 messages.
- For 5 messages: number 2) with two probabilities in the high uncertainty range (around 0.5) has greater H than example 1) with no probabilities in the high uncertainty range, and is also greater than 3) with uniform probabilities of 0.2.
- For 52 messages: number 7) with a single p=0.4 has greater H than examples 5) or 6) with no probabilities in the high uncertainty range.
- Uniform probabilities do not necessarily yield maximum H for multiple messages (as in the 52 message examples)

5. Some applications

5.1 A re-interpretation of Shannon's English alphabet experiment

The following re-examination of Shannon's English alphabet experiment will be used to demonstrate a clear and explicit interpretation of the experimental results compared with the standard existing interpretation. The experiment is generally considered to determine the entropy of an English letter, which is said to be the amount of information in bits that we obtain on the average (i.e. the information gain) when we learn or read a letter of English in a sentence. Shannon seconded his wife in the experiment to guess the hidden next letter, or alternatively a space, after viewing an unfolding string of characters, i.e. only using knowledge of letters that came before. The number of guesses it took until the correct letter or space was identified was then recorded. So over a number of pages of text from a single book a probability distribution across the range of possible guesses from 1 to 27 could be derived, in relation to the total number of guesses. Using the entropy

expression for this discrete probability distribution of number of guesses until correct one, the H value could then be computed for the English alphabet.
Across a number of experiments Shannon found that H varied between 0.8 and 1.3. Subsequent tests, some on a large scale such as by Moradi *et al.* [21], have also determined H to be between 1.0 and 1.6. Thus, Shannon interpreted H to be the *information content* of letters in the English alphabet. If such a probability distribution was to be uniform across the number of possible guesses for each letter or space we would have: $H = \log_2 27 = 4.755$ bits.

However, we can now apply the previous understanding of Shannon entropy as the number of quantised symbols required to identify an incoming probabilistic message. In Shannon's experiment the probabilistic message is the number of guesses to correct letter identification. Then by the above interpretation of entropy, H=2 will enable 2 binary symbols to identify 4 different numbers of guesses. It is important to note that from H values we cannot derive the mean number of guesses for the whole probability distribution which of course can be determined from the distribution itself. And applying our new interpretation to the H range 1.0 to 1.6 derived from the duplication of the experiment by Moradi *et al.* described in [21]:
For H = 1 bit, up to 2 numbers of guesses can be identified (i.e. 1 or 2).
For H = 1.6 bits, up to 3.03 numbers of guesses can be identified (i.e. 1, 2, or 3).

Thus, the H values in those experiments showed that the *average range* of guesses required to pick the correct letter or space is between 1 and 3, over the numerical range of all correct guesses in the experiment (e.g. sometimes 5 or 9 or even 12 guesses for rare letter combinations). This interpretation of H would seem to be more valid than stating that H depicts the amount of information that an English letter provides. The experiments by Moradi *et al.* also showed that H is primarily dependent on the skill of the human subject and complexity of the text, rather than some intrinsic property of English alphabet. It is also important to note that Shannon's derivation of his channel capacity expression [22] does not rely on entropy. The mathematical derivation of the channel capacity formula in that paper was based on the results of Hartley and Nyquist, in addition to some basic waveform mathematics. Thus Shannon entropy simply estimates the number of quantised symbols to represent the average range of guesses, but not the average number of guesses itself. And from this concise meaning it can be appreciated why H does not quantify the amount of uncertainty or information in a probability distribution, but is rather a dependent function of it as is the mean of any probability distribution.

5.2 Valid selection of the log base

In the inductive derivation of Shannon entropy in section 4.2, we showed that the logarithmic base represents the degree of coding symbol quantisation. And because conversion between log bases can be made using a multiplicative constant, ratios of measures will be unaffected, but differences or sums of measures with an arbitrary log base will not be equivalent to measures that use a base matching the degree of coding symbol quantisation. Thus, sums of differences of entropy values will differ in magnitude from true entropy values with a log base that matches the symbol quantisation, by the conversion constant between the true base and value being used. Since the base determines the units of the measure scale (bits, digits etc.) it would then seem prudent to always use a log base in entropy calculations which matches the degree of coding symbol quantisation, which is 2 for binary symbols.

These conclusions directly impact on methods for computing measures of total uncertainty when hybrid forms of uncertainty exist. As previously discussed, many measures that have been proposed to assess the total uncertainty in information have floundered when they tried to combine measures that intrinsically have different units. As yet there seems to be no consensus on what is a robust and valid approach to derive measures of total uncertainty. Some approaches, such as by DeLuca and Termini [23], could also be criticised because they insert non-probabilistic uncertainty measures into the Shannon entropy expression. The problem with doing that is that the entropy expression has no meaning with non-probabilistic uncertainty measures, such as fuzzy set membership grades in [23], because it is a probabilistic expectation based on relative frequencies. There can be little doubt that other kinds of uncertainty besides probabilistic likelihoods exist in many fields where uncertainty measures are applied for reasoning, and that practical methods would be useful for modelling hybrid uncertainty. This author has also proposed [24] one practical approach to modelling hybrid uncertainty which allows hybrid uncertainties to be systematically propagated through mathematical equations.

6. Summary

Shannon stated in his seminal paper that any monotonic function of the number of messages generated by a source can be regarded as a measure of the information being generated. One consequence of using any function of the number of messages and their probabilities is that the meaning of the measure becomes obscure, and consequently is open to many different interpretations. For example, Shannon also proposed the term "measure of choice" as an alternative to the measure of information for his entropy expression, i.e. more messages means more to select from or identify. Thus the terms Shannon suggested for his expression are: measure of information, measure of uncertainty, and measure of choice. This report has shown that Shannon entropy simply represents the expected number of quantised (e.g. binary) symbols required to discriminate or identify an incoming probabilistic message. This follows directly from the definition of the logarithmic function. Being an expectation based on probabilities, there will be some rare messages with low probabilities that will require more bits to identify them than the number computed by the entropy expression. Applying this meaning to Shannon's English language experiment we also derived a clear interpretation of those results.

In recent years there has been increasing research on measures for higher-order and hybrid uncertainty forms which occur across many application domains, e.g. [25]. As a preliminary step towards developing higher-order and hybrid uncertainty measures it would be beneficial, if not essential, to have a clear understanding of any special probabilistic uncertainty measures that currently exist such as Shannon entropy. Thus, the primary motivation behind this re-examination of Shannon entropy has been to gain a more precise interpretation so that the measure is more effectively applied, especially in Defence applications such as for word sense disambiguation in automated text analysis.

7. References

1. Shannon, C. (1948) A mathematical theory of communication, The Bell System Technical Journal, 27(3), pp. 379-423.
2. Shannon, C. and Weaver, W. (1949) The Mathematical Theory of Communication, Univ. Illinois Press: Urbana, Ill.
3. Seife, C. (2006) Decoding the Universe, Viking: London.
4. Schiffer, M (1991) Shannon's information is not entropy, Physics Letters A, 54(7,8), pp.361-365.
5. Daroczy, (1970) Generalized information measures, Information and Control, 16, pp36-51.
6. Taneja, I. (1984) On characterization of generalised information measures, J. Comb. Information and Systems Science, 9, pp. 169-174.
7. Aczel, J and Daroczy, Z (1975) *On Measures of Information and Their Characterizations*, Academic: New York.
8. Gottinger, H (1975) Lecture Notes on concepts and measures of information, Institute of Mathematical Economics, Universal: Bielefeld.
9. Haynes, K. Phillips, F. and Mohrfield, J. (1980) The entropies: Some roots of ambiguity, Socio-Economic Planning Sciences, 14(3), pp. 137-145.
10. Styre, D. (2000) Insight into entropy, American J. Physics, 68(12), pp. 1090-1096.
11. Kahre, J (2002) The Mathematical Theory of Information, Kluwer: Dordrecht.
12. Klir, G. and Folger, T. (1988) Fuzzy Sets, Uncertainty, and Information, Prentice Hall: New York.
13. Pal, N. and Bezdek, J. (1994) Measuring fuzzy uncertainty, IEEE Trans. Fuzzy Systems, 2(2), pp.107-118.
14. Higashi, M. and Klir, G. (1982) On measures of fuzziness and fuzzy complements, Int. J. General Systems, 8, pp. 169-180.
15. Pal, N. (1999) On quantification of different facets of uncertainty, Fuzzy Sets and Systems, 107, pp. 81-91.
16. Yager, R. (1992) On the specificity of a possibility distribution, Fuzzy Sets and Systems, 50, pp. 279-292.
17. Yager, R. (1983) Entropy and Specificity in a Mathematical Theory of Evidence, Int. J. General Systems, 9, pp. 249-260.
18. Zadeh, L. (2005) Toward a generalized theory of uncertainty, Information Sciences, 172, pp. 1-40.
19. Klir, G. (2004) Generalized information theory: aims, results, and open problems, Reliability Engineering and System Safety, 85(1-3), pp.21-38.
20. Hartley, R. (1928) Transmission of information, The Bell System Technical Journal, 7, pp. 535-563.
21. Moradi, H., Gryzmala-Busse, J., Roberts, J. (1998) Entropy of English Text: Experiments with humans and a machine learning system based on rough sets, Information Sciences, 104,(1-2), pp. 31-47.

22. Shannon, C. (1949) Communication in the presence of noise, Proc. IRE, 37(1), pp. 10-21.
23. DeLuca, A. and Termini, S. (1972) A definition of a non-probabilistic entropy in the setting of fuzzy set theory, *Information and Control*, 20(4), pp.301-312.
24. Warren, L. (2007) On Modelling Hybrid Uncertainty in Information, *DSTO Research Report (DSTO-RR-0325)*.
25. Li, X and Liu , B. (2009) Chance measure for hybrid events with fuzziness and randomness, *Soft Computing*, 13, pp.105-115.

22. Shannon, C. (1949) Communication in the presence of noise, Proc. IRE, 37(1), pp. 10-21.

23. DeLuca, A. and Termini, S. (1972) A definition of a non-probabilistic entropy in the setting of fuzzy set theory, *Information and Control*, 20(4), pp.301-312.

24. Warren, L. (2007) On Modelling Hybrid Uncertainty in Information, *DSTO Research Report (DSTO-RR-0325)*.

25. Li, X and Liu, B. (2009) Chance measure for hybrid events with fuzziness and randomness, *Soft Computing*, 13, pp.105-115.

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION **DOCUMENT CONTROL DATA**	1. DLM/CAVEAT (OF DOCUMENT)
2. TITLE A New Interpretation of the Shannon Entropy Measure	3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION) Document (U) Title (U) Abstract (U)
4. AUTHOR(S) Lewis Warren	5. CORPORATE AUTHOR DSTO Defence Science and Technology Organisation PO Box 1500 Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER DSTO-TN-1395	6b. AR NUMBER AR-016-211	6c. TYPE OF REPORT Technical Note	7. DOCUMENT DATE January 2015

8. FILE NUMBER 2014/1062396/1	9. TASK NUMBER NS 07/418	10. TASK SPONSOR Head Languages, Technologies and Fusion	11. NO. OF PAGES 14	12. NO. OF REFERENCES 25

13. DSTO Publications Repository http://dspace.dsto.defence.gov.au/dspace/	14. RELEASE AUTHORITY Chief, National Security and ISR Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS

Uncertainty measures, Information measures, Entropy

19. ABSTRACT

Although more than sixty years have elapsed since Shannon's seminal information entropy paper the literature reveals that there are divergent opinions of what it actually measures. From its similarity to Boltzman entropy in statistical mechanics, the most common view is that it measures information disorder and uncertainty. Based on an inductive derivation of the expression we propose a new interpretation relating to the use of symbols to uniquely identify probabilistic messages. Applying this new meaning to Shannon's English language experiment we extract a new interpretation of those results. Moreover, the new understanding of Shannon entropy also has the potential to improve the effectiveness of intelligence analysis applications.